Feature-based Joint Planning and Norm Learning in Collaborative Games

Mark K Ho[1] (mark_ho@brown.edu), James MacGlashan[2] (james_macglashan@brown.edu), Amy Greenwald[2] (amy@cs.brown.edu), Michael L. Littman[2] (mlittman@cs.brown.edu), Elizabeth M. Hilliard[2] (betsy@cs.brown.edu), Carl Trimbach[2] (ctrimbac@cs.brown.edu), Stephen Brawner[2] (sbrawner@cs.brown.edu), Joshua B. Tenenbaum[3] (jbt@mit.edu), Max Kleiman-Weiner[3] (maxkw@mit.edu), Joseph L. Austerweil[1] (joseph_austerweil@brown.edu)

[1]Department of Cognitive, Linguistic, and Psychological Sciences, 190 Thayer St., Providence, RI 02912 USA
[2]Computer Science Department, 115 Waterman St., Providence, RI 02912 USA
[3]Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

Abstract

People often use norms to coordinate behavior and accomplish shared goals. But how do people learn and represent norms? Here, we formalize the process by which collaborating individuals (1) reason about group plans during interaction, and (2) use task features to abstractly represent norms. In Experiment 1, we test the assumptions of our model in a gridworld that requires coordination and contrast it with a "best response" model. In Experiment 2, we use our model to test whether group members' joint planning relies more on state features independent of other agents (landmark-based features) or state features determined by the configuration of agents (agent-relative features).

Keywords: joint intentionality; norms; team reasoning; reinforcement learning; features; computational modeling

Introduction

From driving to running institutions like the U.S. Postal Service, groups need to coordinate their behaviors to accomplish shared goals. Key to this is that agents learn and use *norms* to guide individual and collective behavior. But how do people (or more generally, how can any intelligent agent) represent and learn norms?

One approach is to treat coordination as emerging through the ego-centric behavior of individual agents. For instance, norms can emerge when agents have other-regarding or aligned preferences (Binmore, 2010). Other approaches use off-the-shelf algorithms, like Q-learning, to show how under certain reward structures, "socially-blind" learning mechanisms can produce social norms (Sen & Airiau, 2007; Claus & Boutilier, 1998). More sophisticated approaches allow agents to model others and best respond to the predictions of those models. For example, agents can recursively reason about one another in *cognitive hierarchies* (Camerer et al., 2004; Wunder et al., 2011).

These computational approaches generally make two key assumptions: First, norms are modeled as *emergent* behavioral by-products rather than *intended* outcomes of agents' learning mechanisms. Second, the space of possible norms is generally constrained to a small set of singular actions (e.g. *cooperate* or *defect* in Flood and Dresher's Prisoner's Dilemma). As a result, the *representation of a norm* is never distinguished from the low-level actions that instantiate the norm.

Unfortunately, psychological research and intuition raise doubts about applying these assumptions to people. For instance, people take a group perspective when choosing their actions in coordination games using *focal points* (Schelling, 1960; Bardsley et al., 2010). Similarly, norms like "curb your dog" seem to rely on learned abstract representations that are applied flexibly to new situations. With this in mind, we have formulated a computational model that incorporates two novel properties:

(1) Agents reason *as if* they were part of a single agent with joint mental states like beliefs, desires, and plans. For instance, a postal worker does not simply reason in terms of "I-intentions" (e.g. I will bring these letters to this address), but also in terms of "we-intentions" (e.g. I will deliver these letters so *we* can deliver the mail) (Searle, 1995; Bacharach, 2006).

(2) Norms are represented as *joint planning biases* that reflect instructions to perform (or avoid) actions. Following Biccheri (2006), agents both follow these instructions and expect others to as well. Formally, these are feature-based reward functions for when an agent plans actions. This provides a compact representation for norms that enables generalization.

This model represents a first step towards understanding how norms are learned through joint reasoning and represented abstractly, aspects of human norm learning not captured in previous formulations.

To study how people learn norms, we focus on multi-state, multi-round coordination games. In our tasks, payoffs are always shared and depend on multi-state planning to reach individual goals simultaneously. In two experiments, we examine the extent to which our model captures how people learn norms. Experiment 1 compares people's behavior to our Norm-Learning model and a Best-Response model that plans actions optimally according to a learned model of its partner. Experiment 2 uses the Norm-Learning model to examine how people generalize norms across situations and the extent to which they use landmark-based or agent-relative features.

Computational Models

Norms are instructions that individuals follow and expect others to follow. We formalize this notion and describe how a group of norm-following agents can converge on norms in a decentralized manner.

Multi-Agent Decision Making

Markov Decision Processes (MDPs) and Stochastic Games MDPs model single-agent decision making and are defined by the tuple (S, A, T, R): a set of states in the world, S; a set of actions the agent can take, A; transition dynamics, $T(s'|s,a)$, which assign a probability of transitioning to a state $s' \in S$ after an agent takes action $a \in A$ in state $s \in S$; and a reward function $R(s,a,s')$, which returns a real valued reward when transitioning to state s' after the agent has taken action a in state s'. An agent in an MDP has a policy (a mapping from states to actions) $\pi: S \rightarrow A$. An agent's policy relates directly to the value, or expected future discounted reward, of each state: $V(s_0) = E_\pi[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1})]$, where $\gamma \in [0,1]$ is a discount factor specifying the value of immediate rewards relative to temporally distant ones. Here, $\gamma = .95$.

To find an optimal policy, an agent needs to calculate the optimal state ($V^*(s)$) and state-action ($Q^*(s,a)$) value functions. Given these functions, the optimal policy can be derived by taking the action with the highest value: $\pi^*(s) \in \text{argmax}_{a \in A}\, Q(s,a)$ (Sutton & Barto, 1998).

MDPs can be extended to include multiple agents using game theoretic tools (Littman, 1994). A *stochastic game* is defined by the tuple (I, S, A^I, T, R^I), where I is an index set of agents in the environment, S is the set of states; A^I is the set of actions for each of the agents with A^i denoting the action set of agent $i \in I$; $T(s'|s,j)$ defines the task dynamics by specifying transition probabilities given a *joint action*, $j \in \times_i A^i$, of all agents taken in state $s \in S$; and R^I is a set of reward functions for each agent, with $R^i(s,j,s')$ denoting the reward received by agent $i \in I$ when agents in state $s \in S$ take joint action $j \in \times_i A^i$ and the environment transitions to state $s' \in S$.

Because multiple agents with individual reward functions are involved, there is no direct analogue of an 'optimal policy' in stochastic games. Rather, solution concepts can be posited (e.g. Nash equilibria) or a learning mechanism can determine how the multi-agent system converges.

Norm-Learning Model

Norms as reward function biases We assume that norms are instructions that an agent (a) follows, and (b) expects others to follow. More formally, we first represent the instructional content of norms as *reward biases* that cause a group of agents to prefer certain types of actions or states. For example, a norm to "drive on the right" would be represented as a collective preference for states that satisfy that description. This provides a natural, flexible way to represent the content of norms. To simplify the problem, we assume the norm bias is based on a linear combination of state features. Assuming agents have a feature function, Φ, that maps states to feature vectors, the norm bias is represented as:

$$R_{\text{norm}}(s) = \theta^T \Phi(s)$$

where θ^T is the transpose of a feature weighting vector. This allows the model to learn that certain state features (e.g. being on the right) is preferable during joint planning.

Second, we incorporate the motivational influence of norms directly into individual agents' reward functions that are used to calculate a reward-maximizing policy. Formally, for the i-th agent in a community, their total reward function will combine their private reward function and a norm bias:

$$R^i(s) = R^i_{\text{individual}}(s) + R_{\text{norm}}(s).$$

All agents in the community will have the same norm bias, R_{norm}, and know other agents will follow it. Thus, norms are joint reward function biases that agents follow and expect other agents to follow.

Inferring Norms We implement learning norms as *group inverse reinforcement learning* (IRL). In single-agent IRL one observes an agent behaving in an MDP and based on those observations infers the goals or reward function of the agent (Abbeel & Ng, 2004; Baker et al., 2009).

A Norm-Learning agent attempts to infer the norm that a group follows given some history of group interaction. That is, each agent estimates the most likely norm given a history of interaction, $\mathcal{H} = \big((s_0, j_0, s_1), \ldots, (s_{T-1}, j_{T-1}, s_T)\big)$:

$$\hat{R}_{\text{norm}} = \underset{R_{\text{norm}}}{\text{argmax}}\, P(R_{\text{norm}} | \mathcal{H}).$$

Since the norm bias function is a linear weighting of features, this corresponds to finding the most likely weights, $\hat{\theta}$.

Here, we focus on norm learning in collaborative games. That is, we assume that all agents all have the same goal (i.e. have the same $R^i_{\text{individual}}$) but must figure out how to work together to accomplish it. This simplifies inferring the norm bias. Other work should investigate how norm learning interacts in competitive scenarios (e.g. see Kleiman-Weiner et al. in this year's proceedings).

Features for Learning Norms Our representation of norms and implementation of norm learning depends on the features available to individuals in the group. The specific features are important for several reasons. First, to converge on a norm, individuals must have sufficiently similar features available to them to determine which norm the group uses. Second, features must be sufficiently expressive to allow individuals to pin down the norm that they collectively use to solve a task. Third, learning norms in terms of features allows generalization to novel situations. Without a concise, abstract representation of a norm, people would not be able to apply a learned norm to a new context and would need to learn an appropriate norm from scratch. For the tasks in the experiments reported, we describe which types of features are used for constructing norms.

We consider two types of features: *landmark-based features*, such as "Agent X is north of its goal", and *agent-relative features*, such as "Agent X is north of Agent Y". These two types were chosen because the former are an 'asocial' representation, while the latter explicitly involve social others. Moreover, they lead to different predictions in the tasks we use.

Best-Response Model

Best-response agents individually plan using a model of other agents. This means that instead of reasoning about a joint-policy directly, an agent i uses a predictive model of another agent j's policy, $\hat{\pi}_j$, to predict what j will do in a certain state. That is, an agent i will construct a transition function that includes predictions about the behavior of the other agent, $\tilde{T}_i(s'|s,a_i) = T\left(s'\middle|s,\left(a_i,\hat{\pi}(s)\right)\right)$.

Here, we use a level-1 cognitive hierarchy planner as our Best-Response model. It models its partner's behavior directly by counting its partner's actions and decaying past counts by a parameter δ. Additionally, to accelerate learning, the model assigns a pseudocount, α, to joint states in which the partner's location on the grid is the same[1]. Although we could have modeled a higher level of reasoning (e.g. best responding to a level-1 planner) we did not for two reasons. First, previous experimental work has shown that people do not typically reason beyond one or two levels (Camerer et al., 2004). Second, in non-competitive contexts, strategies often converge at higher levels in the cognitive hierarchy, and even level-1 reasoning provides a good estimate of this converged behavior (Bardsley et al., 2010).

Experiment 1: Hallway Task

Task Description

To test whether people best respond or learn norms, we designed the 2-person Hallway task shown in Figure 1. Two agents (circles) start at opposite ends of a 5x3 grid and on each turn simultaneously move up, down, left, right, or wait. The two agents cannot enter the same state or immediately switch positions – if they attempt this, then they collide and remain in the same location as in the previous time-step. Each agent has its own goal tile, indicated by a matching color, and the two agents start the task on one another's goals. Whenever ***either*** agent enters its own goal state, the round ends. However, to succeed in a round (and in the human case win a bonus), the agents must ***simultaneously*** enter their respective goals. This necessitates collaboration.

At the beginning of a round, each agent is exactly 4 tiles away from its goal. But they cannot both take a direct route to their goals without colliding. Rather, they must choose a series of actions that enables them to "break the symmetry"

[1] For example, if the agent is at (1,1) and the partner is at (0,0), then the partner's behavior will be generalized to other joint states where it is at (0,0). In this paper $\delta = 0.5$, $\alpha = 0.5$.

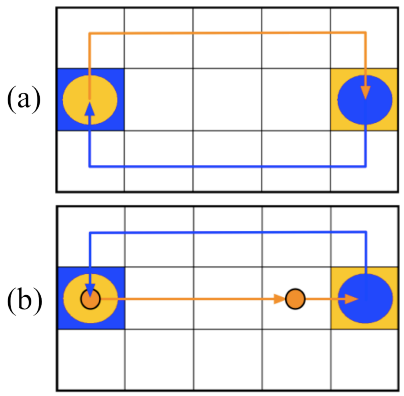

Figure 1: Hallway Task examples where (a) agents pass on top and bottom rows, and (b) they pass on middle and top rows. Smaller circles indicate the agent waited a step.

and pass one another. Critically, at least one of the two agents has to deviate from the center row of the grid and return to the center for the two to successfully complete the task. The other agent will either have to do the same, but on a different row, or wait two time-steps for the other agent. Figure 1 displays two joint plans that successfully complete the task in the minimal number of steps (6). Note though that there are many other possible joint actions that the two agents can take to pass one another.

In a given round, we can consider the row that each player is on when the two pass. Each player can be either on the *top*, *bottom*, or *center* row when attempting to move closer to their own goal. Clearly, successful passing requires that the two agents be on different rows while attempting to pass. Figure 2 visualizes this as an outcome matrix. Executing a successful joint policy, defined as both agents reaching their goal in the minimal number of steps, requires that both agents select different passing rows.

		Player 2's Passing Row		
		Top	Center	Bottom
Player 1's Passing Row	Top	*Fail*	*Success*	*Success*
	Center	*Success*	*Fail*	*Success*
	Bottom	*Success*	*Success*	*Fail*

Figure 2: Matrix representing passing success as a function of each player's row in the gridworld. Note that "Success" means the game was solved optimally.

Model Simulations

Best-Response Suppose two Best-Response agents succeed where player 1 passes through the center and player 2 passes along the top ({*center, top, success*}). Having observed player 2's behavior, player 1's predicts that player 2 will again choose *top*. From player 1's perspective, it is equally optimal to choose *center* as it is to choose *bottom*. However, if player 2 reasons similarly about player 1, then player 2 will treat *top* or *bottom* as equally optimal. If the players choose their respective pairs of equally optimal actions at

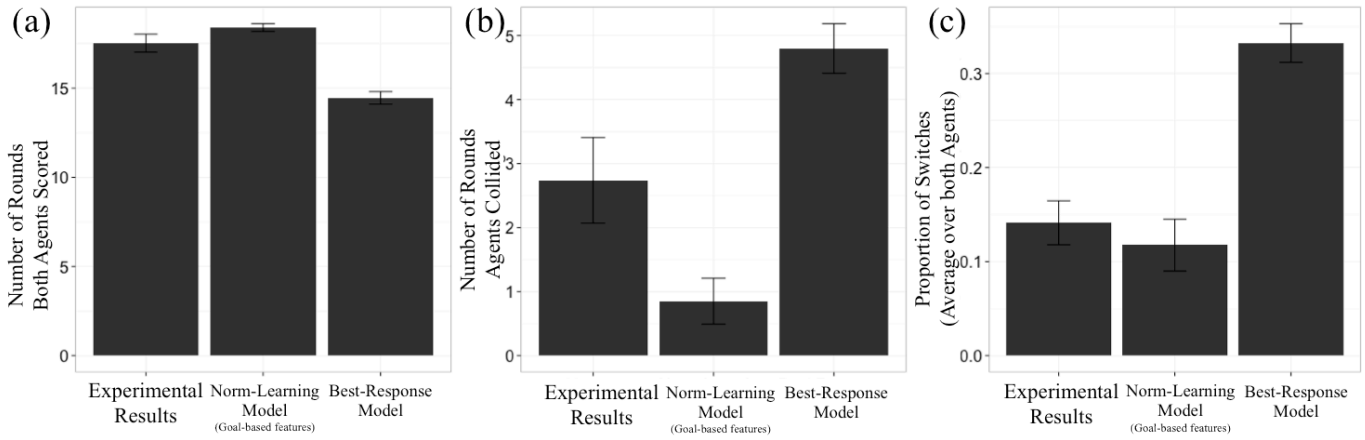

Figure 3: Human (experimental), Norm-Learning (simulation), and Best-Response (simulation) results for (a) number of rounds (out of 20) in which both agents scored, (b) number of rounds in which agents collided at least once, and (c) proportion of rounds in which agents switched their strategy from the previous round (averaged over both agents). These results suggest that human collaboration relies on jointly learned norms rather than best-responding to one's partner.

random (i.e. 0.5 each), then they will stay at {*center, top, success*} with .25 probability, switch to the {*bottom, top, success*} or {*center, bottom, success*} cells with 0.5 probability, or switch to {*bottom, bottom, fail}* with .25 probability. In our implementation, the probabilities differ from this ideal due to the decaying memory. Nonetheless, this illustrates the central prediction of best-response decision making in this collaborative game: that there will be high row-passing switching as well as a moderate amount of collisions from agents simultaneously switching.

Note also that the memoryless, mixed-strategy Nash equilibrium is itself a type of best-response solution concept. In this particular coordination game it is $\langle 1/3, 1/3, 1/3 \rangle$, which leads to an even higher proportion of collisions – $1/3$ – as well as switching – $2/3$.

Norm-Learning In the Norm-Learning model, two agents observe the same history of interaction, and use this to infer the most likely norm that a hypothetical collective agent is using. By using their shared observations and reasoning processes to deduce the most likely norm that they as a group have, they converge on and stay with a particular norm. For the Hallway task, we use a set of *landmark-based features* to define the space of norm reward biases. Specifically, for each of the two agents, we represent which row they are on relative to the row that their goal is on: above (top), on (center), or below their goal's row (bottom). This gives us a total of 6 binary goal-based features.

Unlike the Best-Response model, the Norm-Learning model predicts that people will stick with a combination of rows when performing the task. That is, in dyads that collaborate successfully, participants will not change which row they pass on and there will be few, if any, collisions.

Experiment

Design and Procedure We recruited 50 MTurk participants (25 dyads). They signed a consent form and then completed demo tasks that familiarized them with the grid game interface and task dynamics. They received a $2.00 base payment and an additional $0.10 bonus when they simultaneously reached their goals. Afterwards, each participant completed a post-task survey that included questions about the task and demographics. One dyad was excluded from analysis due to missing data.

For the simulations, agent dyads played 20 rounds and only learned at the end of each round. The Best-Response model updated its model of its partner based on the play the previous round, while the Norm-Learning model updated its distribution over possible norm biases. To simplify inference, we considered the space of feature weights to be $\theta \in \{-1,0,1\}^{f_n}$.

Results and Discussion Participants reported the task being relatively easy where 1 = Very Difficult to 7 = Very Easy (Mean = 5.67; SE = .19). Additionally, participants reported that they were skilled at the task on a scale from 1 = Very Bad to 7 = Very Good (Mean = 5.64; SE = .18).

The dyads were successful at collaborating on the task and winning the bonus. For our analysis, we focused on dyads that scored more than half of the rounds (23 of 24 dyads). These dyads, on average, jointly scored 17.5 out of 20 rounds (SE = 0.50) and jointly scored in the minimum number of steps possible (6) 15.22 out of 20 rounds (SE = 0.85). Human rounds scored did not differ from the Norm-Learning model ($t(35.307) = 1.82$, $p = 0.07$) but did differ from the Best-Response model ($t(38.0) = 4.98$, $p < .001$) (Figure 3a). However, direct comparison by scoring is difficult since the simulations update only once a round completes. This leads the Best-Response model to potentially collide indefinitely and never reach its goal.

Overall, the experimental results resemble the predictions of the Norm-Learning model over the Best-Response model. To compare human behavior in the collaborative Hallway task to the models, we focused on two measures: the number of rounds in which the agents collided at least once,

and the proportion of rounds in which agents switched their strategy from the previous round.

Figure 3b plots the number of rounds in which at least one collision occurred for a dyad. Human collisions fell directly in between the Norm-Learning model and Best-Response model and significantly differed from the numbers in both models (Best-Response: t(34.7) = -2.67, p = 0.01; Norm-Learning: t(33.2) = 2.49, p = 0.02). While this fails to distinguish between the two idealized models presented, it is consistent with a noisy or imperfect norm learning process.

To determine what strategy human and simulated agents used in a given round, we used the following heuristic: the first time a dyad collided in a round, the respective rows of the agents were coded as their strategies (note that agent locations after a collision remain the same as before). If a dyad never collided in a round we looked at the first step in which they were in the same column or switched columns -- i.e. the step at which they passed one another. Their respective rows *after* passing were coded as their strategy.

The amount of trial-to-trial strategy-switching provides a critical contrast between the Norm-Learning model and Best-Response model. Individuals in the human dyads switched their strategies rarely (14% of rounds) (Figure 3c). This closely matches the predictions of the Norm-Learning model ($t(38.8) = 0.66$, $p = .51$) and deviates significantly from the Best-Response model ($t(40.9) = -6.14$, $p < .001$). A related measure of joint strategy diversity tells us about global rather than local variability. To measure this, we calculated the entropy of the frequency distributions of joint strategies. Although the joint strategy entropy for humans significantly differed from both models (Norm-Learning: t(36.6) = 2.81, p < .001; Best-Response: t(36.2) = -5.55, p < .001), human entropy was closer to Norm-Learner entropy (Human: M = 0.68, SE = 0.10; Norm-Learning: M = 0.34, SE = 0.06; Best-Response: M = 1.34, SE = 0.06). Overall then, people displayed behavior more consistent with learning a jointly understood norm rather than best responding to their partner's behavior.[2]

Experiment 2: Feature-Based Norm Generalization

In our model, feature-based norms allow agents to converge on a joint strategy quickly, but they also enable individuals to generalize norms to new contexts. Here, we present the predictions of two sets of landmark-based features and one set of agent-relative features. We show that human norm learning better resembles landmark-based features assuming people learn norms like the Norm-Learning model.

Task Description

We designed the Courtyard tasks shown in Figure 4 to test norm generalization. In Direct Courtyard, agents can immediately move towards their respective goals through entrances. In Indirect Courtyard, agents must first move upwards towards entrances to cross. But both versions require agents to devise a way to pass. In our simulations and experiment, agents first play 10 rounds of Direct Courtyard, followed by 10 rounds of Indirect Courtyard.

If the agents have no means of generalizing learned joint plans, then they will have to find a new one when moving from Direct to Indirect Courtyard. On the other hand, if they learned a norm as a feature-based reward bias, then they can use it to guide their strategy choice on Indirect Courtyard.

Models

We tested three sets of binary feature-based reward biases: goal-relative features – is agent X's current row above, below, or the same as the goal row? (f_n = 2 agents x 3 features = 6 total features) –, entrance-relative features – is agent X's current row above, below, or the same as the entrances? (f_n = 6) –, and agent-relative features – is agent X above, below, or on the same row as agent Y? (f_n = 3). 40 simulations of each model were run.

Experiment

Procedure 90 MTurkers (45 dyads) participated. 3 dyads were excluded due to technical error. The procedure matched Experiment 1 except participants played 2 games.

Results and Discussion To determine whether and how participants generalized from Direct to Indirect Courtyard, we analyzed individual strategies within dominant joint strategies. Each round, we identified the strategy used with the same heuristic as in Experiment 1. For each dyad, we then identified the most frequent strategies in the Direct Courtyard and Indirect Courtyard phases. We then identified each agent's individual strategy in the two phases. Figure 5a shows individual strategy counts in the two phases. Note that if people were forming new joint plans from scratch, the strategies in the two phases would be uncorrelated. This was not the case ($\chi^2(4) = 27.45$, $p < .001$).

To explain the systematicity in how joint plans were generalized to the Indirect Courtyard, we compared this distribution to the simulation results of our three models (Figure 5b, 5c, and 5d). A visual analysis of these tables suggested that the participant results reflect a mixture of goal-relative and entrance-relative features. We confirmed

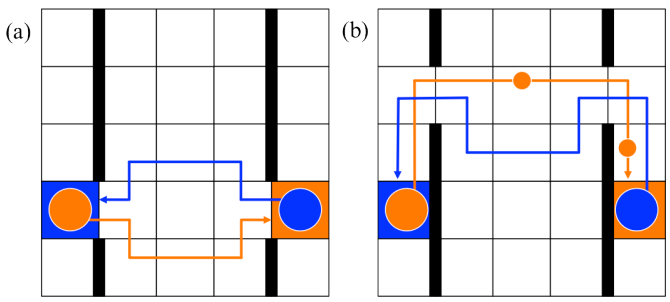

Figure 4: (a) Direct and (b) Indirect Courtyard Tasks with example optimal joint plans. The bold lines are walls that agents cannot pass through.

[2] We assume that best responders do not default to the previous choice if indifferent between options. Future studies will need to rule this out in people.

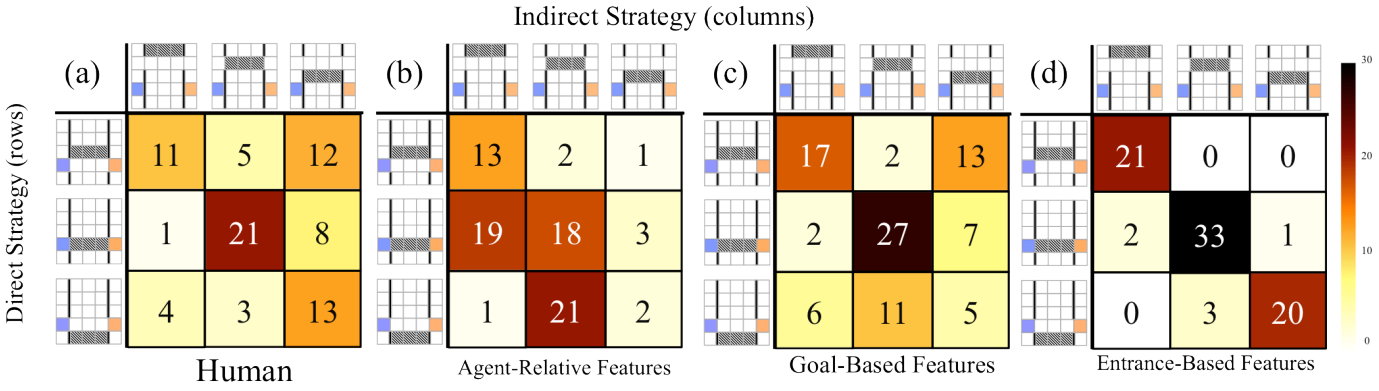

Figure 5: Individual agents' most frequent strategies (counts) by Direct and Indirect Courtyard phases for (a) experimental participants, (b) our Norm-Learning agent with Agent-Relative features, (c) Goal-Based features, and (d) Entrance-Based features. The distribution indicates which row an agent is likely to take in the Indirect Courtyard grid given what row was taken in the Direct Courtyard. Human results are best explained as a mixture of the two landmark-based feature sets (Goal-based and Entrance-based; see text). Grayed out rows in the gridworlds indicate the most frequent individual strategy.

this by calculating maximum likelihood mixture values for the three models: $p_{\text{Entrance}} = 0.21$, $p_{\text{Goal}} = 0.79$, and $p_{\text{Agent}} = 0.0$. Thus, participants tended to generalize norms using landmark-based rather than agent-relative features.

Conclusion

Here, we have presented a novel model of norm learning based on inferring joint reward biases. We compared the predictions of this model to those of a best response model in the Hallway task, and used the same model to show that people use landmark-based rather than agent-relative features to generalize norms across the two Courtyard tasks.

A central aspect of human sociality is engaging in shared intentions and joint plans with others (Searle, 1995). Using the formal tools of multi-agent MDPs, we are able to make quantitative predictions about how collaborating individuals represent and reason about themselves as part of a larger entity. Future work should explore how individuals learn norms over particular types of features, what happens when agents' feature sets differ from one another, and how learned norms interact in competitive scenarios.

Acknowledgments

This material is based upon work supported by the NSF GRF under Grant No. (DGE-1058262).

References

Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In C. Brodley (Ed.), *Proc. of twenty-first international conference on machine learning* (p. 1). New York, NY: ACM.

Bacharach, M., Gold, N., & Sugden, R. (2006). *Beyond individual choice: teams and frames in game theory*. Princeton University Press.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Bardsley, N., Mehta, J., Starmer, C., & Sugden, R. (2010). Explaining Focal Points: Cognitive Hierarchy Theory versus Team Reasoning*. *The Economic Journal*, *120*(543), 40–79.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Binmore, K. (2010). Social norms or social preferences? *Mind & Society*, *9*(2), 139–157.

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 861–898.

Claus, C., & Boutilier, C. (1998). The dynamics of reinforcement learning in cooperative multiagent systems. In J. Mostow, C. Rich, B. Buchanan (Eds.) *Proc. of the 15th national conf. on artificial intelligence* (pp. 746-752).

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In W. Cohen (Ed.), *Proc. of 11th international conf. on machine learning* (pp. 157–163). San Francisco, CA: Morgan Kaufmann.

Schelling, T. C. (1980). *The strategy of conflict*. Harvard university press.

Searle, J. R. (1995). *The construction of social reality*. Simon and Schuster.

Sen, S., & Airiau, S. (2007). Emergence of Norms Through Social Learning. In R. Sangal, H. Mehta, R. K. Bagga (Eds.), *Proc. of the 20th international joint conference on artificial intelligence* (pp. 1507–1512). San Francisco, CA: Morgan Kaufmann.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT press.

Wunder, M., Kaisers, M., Yaros, J. R., & Littman, M. (2011). Using Iterated Reasoning to Predict Opponent Strategies. In L. Sonenberg, P. Stone (Eds.), *10th international conference on autonomous agents and multiagent systems* (Vol 2, pp. 593–600). Richland, SC: IFAAMS.